SiENet: Siamese Expansion Network for Image Extrapolation

Xiaofeng Zhang, Feng Chen, Cailing Wang, Ming Tao, and Guo-Ping Jiang, *Senior Member, IEEE*

***Abstract*—Different from image inpainting, image outpainting has relatively less context in the image center to capture and more content at the image border to predict. Therefore, classical encoder-decoder pipeline of existing methods may not predict the outstretched unknown content perfectly. In this paper, a novel two-stage siamese adversarial model for image extrapolation, named Siamese Expansion Network (SiENet) is proposed. Specifically, in two stages, a novel border sensitive convolution named adaptive filling convolution is designed for allowing encoder to predict the unknown content, alleviating the burden of decoder. Besides, to introduce prior knowledge to network and reinforce the inferring ability of encoder, siamese adversarial mechanism is designed to enable our network to model the distribution of covered long range feature as that of uncovered image feature. The results on four datasets has demonstrated that our method outperforms existing state-of-the-arts and could produce realistic results. Our code is released on https://github.com/nanjingxiaobawang/SieNet-Image-extrapolation.**

***Index Terms*—Two-stage GAN, Adaptive filling convolution, Siamese adversarial mechanism, Image outpainting.**

I. Introduction

IMAGE extrapolation, is to generate new contents beyond the original boundaries of a given image. Even belonging to the general image painting task as inpainting [1]–[6], image outpainting [7] has its special characteristics. We rethink this task from three aspects. First, compared with inpainting, image extrapolation would rely on relative less context to infer much larger unknown content. Second, the content to be inferred locates outside the given image in image outpainting. Therefore, existing expertise of inpainting is not suitable to be applied into this task directly. Third, existing methods of image extrapolation generally employ classical encoder-decoder (ED) structure which forces the encoder to capture the global and local features and allows decoder to recover these features to desired resolution. Namely, all the burden of inferring is endured by the decoder. Yet, the input of outpainting doesn't have such abundant features, which makes the capturing ability of encoder weak and the inferring burden of decoder heavy.

Due to the requirement of outpainting task, a suitable framework is necessary to address the task. However, existing methods which follow the expertise of inpainting still suffer from the incompatibility. [8] proposed the specific spatial expansion module and boundary reasoning module for the requirement of connecting marginal unknown content and inner context. Besides, heavy burden of prediction on decoder may result in the poor performance of generating realistic images. [9] applied DCGAN [10] which mainly focuses on the prediction ability of decoder. Even if this kind of methods could predict the missing parts on both sides of the image, this design would ignore coherence of semantic and content information.

In this letter, to design task-specific structure, we propose a novel two-stage siamese adversarial model for image extrapolation, named Siamese Expansion Network (SiENet). In our SiENet, a boundary-sensitive convolution, named adaptive filling convolution, is proposed to automatically infer the features of surrounding pixels outside known content with balance of smoothness and characteristic. This adaptive filling convolution is inserted to the encoder of two-stage network, activating the sensitivity of encoder for border features. Therefore, encoder could infer the unknown content and the inferring burden of decoder could be alleviated.

In addition, the siamese adversarial mechanism is designed to introduce prior knowledge into network and adjust the inferring burden of each part. In the joint training of two stages, ground truth and covered image are fed into network to calculate the siamese loss. Siamese loss encourages the features of them in the subspace to be similar, leading to reinforced predicting ability of long range encoder. Besides, a adversarial discriminator is designed in each stage to push the global generator to generate realistic prediction. Thus, the whole inferring burden of the network is reasonably allocated.

Our contributions can be summarized as follows:

- We design a novel two-stage siamese adversarial network for image extrapolation. Our SiENet, a task-specific pipeline, could regulate the inferring burden of each part and introduce prior knowledge into network legitimately.
- We propose an adaptive filling convolution to concentrate on inferring pixels in unknown area. By inserting this

Manuscript received July 11, 2020; revised August 20, 2020; accepted August 21, 2020. Date of publication August 26, 2020; date of current version September 23, 2020. The work was supported in part by the National Natural Science Foundation of China under Project 61871445 and in part by the Nanjing University of Posts and Telecommunications General School underProject NY22057. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Wei Li. *(Xiaofeng Zhang and Feng Chen contributed equally to this work.) (Corresponding author: Cailing Wang.)*

Xiaofeng Zhang, Cailing Wang, Ming Tao, and Guo-Ping Jiang are with the College of Automation and College of Artifificial Intelligence, Nanjing University of Posts and Telecommunications, Nanjing 210049, China (e-mail: SemiZxf@163.com; wangcl@njupt.edu.cn; mingtao2000@126.com; jianggp@njupt.edu.cn).

Feng Chen is with the School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210049, China (e-mail: chenfeng1271@gmail.com).

This letter has supplementary downloadable material available at https://ieeexplore.ieee.org, provided by the authors.

Digital Object Identifier 10.1109/LSP.2020.3019705

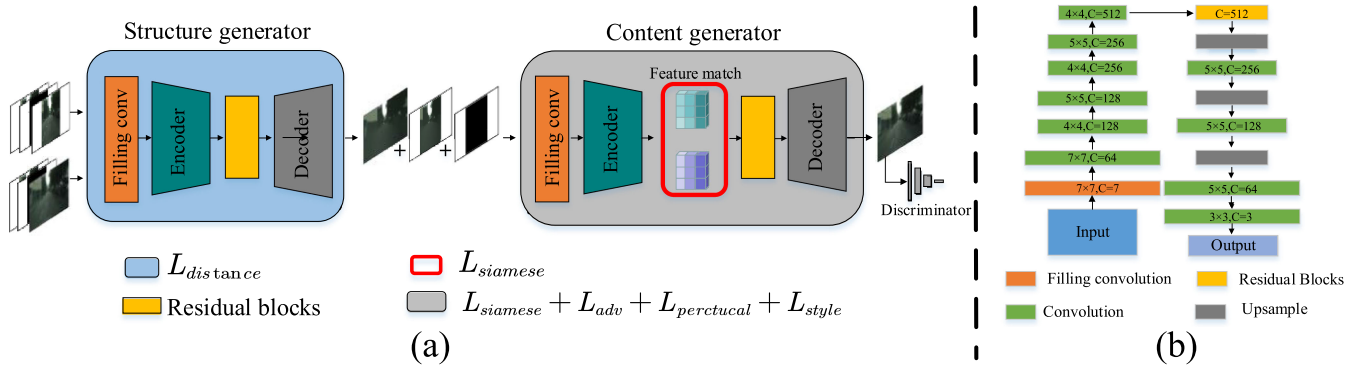

Fig. 1. Illustration of network architecture. (a) is the joint training pipeline of our network. (b) is the detailed structure of structure generator, which is similar to that of content generator.

convolution into encoder, encoder could be endowed powerful inferring ability.

- Our method achieves promising performance on four datasets and outperforms existing state-of-the-arts on four datasets.

II. Methodology

Our SiENet is a two-stage network, and each stage has specific generator and discriminator, as shown in Fig. 1(a). The details of our method are provided as follows.

A. Framework Design

Our SiENet is a coarse-to-fine two-stage network which is composed by a structure generator and a content generator. The whole training of SiENet has three steps: (1) solely train the structure generator; (2) solely train the content generator; (3) jointly train both structure and content generators. At the first step, a covered image $X \in \mathbb{R}^{h\times w\times c}$, an extension filling map $M \in \mathbb{R}^{h\times w\times 1}$ which indicates the known and unknown area, and a covered smooth structure $S \in \mathbb{R}^{h\times w\times c}$ are combined to act as the input of structure generator where w is actual width of covered image. To generate the prediction of uncovered structure where S' and S^{gt} denote the prediction and ground truth of structure respectively. The smooth structure S^{gt} is generated by algorithms [13], [14]. At the second step, we use ground truth structure S^{gt} and covered image X as input of content generator to predict the fully uncovered image $\hat{Y}$. At the third step, as shown in Fig. 1(a), the structure generator and content generator are connected sequentially. The input of the whole network is the same as the first step, however, the input of content generator is the predicted structure S', instead of the ground truth structure S^{gt}, and the covered image X.

As illustrated in Fig. 1(b), the encoder of structure generator downsamples the input with 8× scale. Then to capture multi-scale information two residual blocks are designed to further proceed the output of encoder. Finally, by three nearest-neighbor interpolation, the output of residual blocks is upsampled to desired resolution. The detailed structure of content generator is similar to that of structure generator, except adding two residual blocks before last two upsampling operations. The discriminator D follows the structure and protocol of that of BicycleGAN [15], [16].

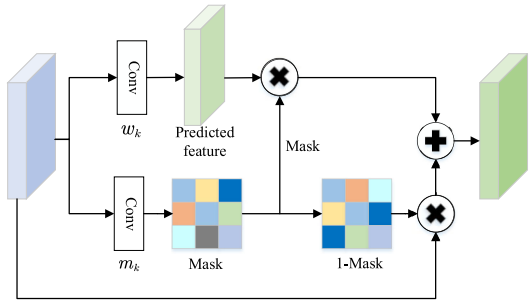

Fig. 2. Scheme of adaptive filling convolution.

B. Adaptive Filling Convolution

To endue encoder with predicting ability, we propose a boundary sensitive convolution, i.e., adaptive filling convolution. This convolution is composed by a predicting function for inferring unknown content and a balance mask for aggregating known content and predicted content. Given a convolutional kernel with K sampling locations (e.g., $K = 9$ in a 3×3 convolution), let w_k and $m_k \in [0, 1]$ denote the weight of predicting function and mask scalar for k-th position. $x(p)$ and $y(p)$ represent the feature map at location p from input and output respectively. Therefore, the filling convolution could be formulated as follows:

$$y(p_0) = \sum_{k=1}^{K} [w_k \cdot m_k \cdot x(p_0 + p_k) + (1 - m_k) \cdot x(p_0 + p_k)], \tag{1}$$

where p_0 refers to arbitrary location in output feature and p_k enumerates K sampling locations in kernel grid. $w_k \cdot x(p_0 + p_k)$ could estimate the unknown content, meanwhile m_k and $1 - m_k$ are multiplied by corresponding predicted content and known content of the same location for keeping smoothness. As the scheme in practical learning shown in Fig. 2, two separate convolutions model predicting weight w_k and mask scalar m_k respectively where m_k is defaulted to 1 to avoid initial over-confidence of maintaining known content. When kernel center locates just outside known content, only scant real pixels would join the patch calculation. The convolution for w_k would predict it by known information, and another skip connection from input to output would keep characteristic value to avoid high variance. Besides, mask scalar m_k works as an adaptive balance

factor of these two branches, to guarantee the smoothness and characteristic of prediction. We insert a 7×7 filling convolution into the bottleneck of encoder, therefore, encoder could possess the ability of capturing context and the ability of predicting unknown content.

C. Siamese Adversarial Mechanism

Classical adversarial mechanism could constrain the whole generator toward the real case. However, this constraint is implicit to each part of generator. Especially in two-stage GAN [17]–[19] for image outpainting, long range encoding may lead to insufficient inferring ability for decoder. Considering the predicting ability of encoder brought by filling convolution, we further add explicit constraint in encoder to push the features of covered and uncovered image to be common. Namely, prior knowledge of uncovered image is learned by encoder. In addition, we also take advantages of adversarial mechanism in our model to constrain the generated results to be consistent with ground truth [20], [21]. Through these two thoughtful constraints brought by our siamese adversarial mechanism, the predicting burden is well regulated.

Let G_{strut} denotes structure generator and E_{cont} denotes encoder of content generator, $I = [X, M, S]$ is the input of the whole network. The output F of long range encoder $E_{cont}(G_{strut}())$ could be formulated as:

$$F = E_{cont}(concat(X, M, G_{strut}(I))) \tag{2}$$

Therefore, let superscripts $'$ and ^{gt} denote whom is generated by covered input and ground truth input respectively, the siamese loss $L_{siamese}$ is $\ell 2$ loss of F' and F^{gt}:

$$L_{siamese} = ||F' - F^{gt}||_2 \tag{3}$$

In order to generate realistic results, we introduce adversarial loss L_{adv} in the structure generator:

$$L_{adv} = \mathbb{E}[\log(1 - D(\widehat{Y})] + \mathbb{E}\left[\log D(Y)\right], \tag{4}$$

where $\widehat{Y}$ and Y are the predicted uncovered image and the ground truth image.

D. Loss Function

We introduce $L_{distance}$ loss to predict the distance between the generated structure S' and ground truth structure S^{gt}:

$$L_{distance} = \left\| S' - S^{gt} \right\|_1 \tag{5}$$

Besides, the perceptual loss $L_{perctual}$ and the style loss L_{style} [22] are applied into our network. $L_{perctual}$ is defined as:

$$L_{perctual} = \mathbb{E}\left[\sum_i \frac{1}{N_i} \left\| \phi_i(Y) - \phi_i\left(\widehat{Y}\right) \right\|_1 \right], \tag{6}$$

ϕ_i is the activation map of the i-th layer of the pre-trained network. In our work, ϕ_i is activation map from the layers of relu1-1, relu2-1, relu3-1, relu4-1 and relu5-1 of the ImageNet pre-trained VGG-19. These activation maps are also used to calculate the style loss to measure the covariance between the activation maps. Given a feature map size $\mathrm{C}_j * \mathrm{H}_j * \mathrm{W}_j$, the style loss L_{style} is calculated as follows.

$$L_{style} = \mathbb{E}\left[\left\| G^{\phi}\left(\widehat{Y}\right) - G^{\phi}(Y) \right\|_1 \right], \tag{7}$$

G^{ϕ} is the the reconstructed Gram metric based on activation map ϕ_j. Totally, the overall loss L_{total} is:

$$\begin{aligned} L_{total} = {} & \lambda_{dist} L_{distance} + \lambda_{adv} L_{adv} + \lambda_p L_{perctual} \\ & + \lambda_s L_{style} + \lambda_{sie} L_{siamese}, \end{aligned} \tag{8}$$

where λ_{dist}, λ_{adv}, λ_p, λ_s and λ_{sie} are set 5, 1, 0.1, 250 and 1 respectively.

TABLE I
ABLATION EXPERIMENTS OF FILLING CONVOLUTION AND SIAMESE ADVERSARIAL MECHANISM ON CITYSCAPES. 'FCONV' AND 'SAM' DENOTE FILLING CONVOLUTION AND SIAMESE ADVERSARIAL MECHANISM RESPECTIVELY

FConv	SAM	SSIM	PSNR
		0.6896	23.2539
√		0.7254	23.5987
	√	0.7647	24.4578
√	√	0.7832	24.8213

Fig. 3. Ablation study of effectiveness of filling convolution and siamese adversarial mechanism on Cityscapes and beach.

III. EXPERIMENTS

A. Implementation Details and Configurations

Our method is flexible for two-direction and single-direction outpainting. For two-direction outpainting, our method is evaluated on three datasets, i.e., Cityscapes [23], paris street-view [2] and beach [9]. Single-direction evaluation is made on Scenery dataset [24]. Besides, we take structural similarity (SSIM), peak signal-to-noise ratio (PSNR) as our evaluation metrics.

In training and inference, the images are resized to 256×256. We train our model using Adam optimizer [25] with learning rate 0.0001, beta1 0 and beta2 0.999. The batch size is set to 8 for paris-street view and beach, 2 for Cityscapes, and 16 for Scenery. The total iteration of joint training is 10^6 for four datasets. Our method has 2 fps at inference time and 56.48 M for inference memory.

TABLE II
PERFORMANCE COMPARISON OF IMAGE TWO-DIRECTION AND SINGLE-DIRECTION OUTPAINTING. THE TWO-DIRECTION OUTPAINTING IS EVALUATED ON BEACH, PARIS-STREET VIEW AND CITYSCAPES, AND SINGLE-DIRECTION OUTPAINTING IS EVALUATED ON SCENERY

	SSIM				PSNR			
Methods \ Datasets	Beach	Paris	Cityscapes	Scenery	Beach	Paris	Cityscapes	Scenery
Image-Outpainting	0.3385	0.6312	0.7135	-	14.6256	19.7400	23.0321	-
Outpainting-srn	0.5137	0.6401	0.7764	-	18.2211	19.2734	23.5927	-
Edge-Connect	0.6373	0.6342	0.7454	-	19.8372	22.1736	24.1413	-
NS-Outpainting	-	-	-	0.6763	-	-	-	17.0267
Our method	**0.6463**	**0.6428**	**0.7832**	**0.8557**	**20.7965**	**23.9794**	**24.8213**	**31.7686**

Fig. 4. Quantitative two-direction comparisons of our method with our state-of-the-arts on Cityscapes.

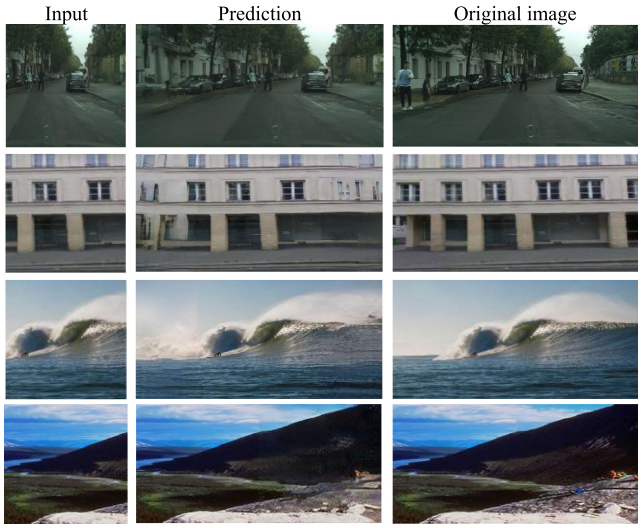

Fig. 5. The visual performance of our method on four dataset. The front three rows represent two-direction outpainting results. The last row is single-direction outpainting result.

B. Ablation Study on SiENet

To validate the effectiveness of filling convolution and siamese adversarial mechanism in our network, we make the ablation experiment in Table I. Only using filling convolution or siamese adversarial mechanism could get a significant improvement over classical encoder-decoder. And the combination of them could achieve the best performance, indicating the regulation of predicting burden is beneficial for the task. We also visualize the results of street and nature scenery to measure the quantitative capability of them. As shown in Fig. 3, the combination of these proposed two parts could recover the details and generate realistic results in easy nature and complicated street cases.

C. Comparisons With Existing Methods

Our method is compared with existing methods including Image-Outpainting [9], Outpainting-srn [8], Edge-connect [26] and NS-outpainting [24]. As shown in Tab. II, for two-direction outpainting, our method outperforms Image-Outpainting [9], Outpainting-srn [8], Edge-connect [26], and achieves state-of-the-art performance in three datasets. For single-direction, the great margin of performance between our method and NS-outpainting [24] indicates the superiority and generality of our SiENet. Besides, we also provide visual comparison of various methods in Fig. 4 and visual performance of our method on four datasets in Fig. 5. Notably, the images generated by our method are similar to the ground truth than that of existing methods.

IV. CONCLUSION

A novel end-to-end model named SiENet is proposed in this paper for image extrapolation. To regulate the heavy predicting burden of decoder legitimately, we first propose adaptive filling convolution to endow encoder with predicting ability. Then we introduce siamese adversarial mechanism in long range encoder to allow encoder to learn prior knowledge of real image and reinforce the inferring ability of encoder. Our method has achieved promising performance on four datasets and outperforms existing state-of-the-arts.

REFERENCES

[1] C. Barnes *et al.*, "PatchMatch: A randomized correspondence algorithm for structural image editing," *ACM Trans. Graph.*, vol. 28, no. 3, pp. 24–33, 2009.

[2] D. Pathak *et al.*, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2536-2544.

[3] J. Yu *et al.*, "Generative image inpainting with contextual attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5505–5514.

[4] C. Yang *et al.*, "High-resolution image inpainting using multi-scale neural patch synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6721–6729.

[5] J. H. Cho *et al.*, "Hole filling method for depth image based rendering based on boundary decision," *IEEE Signal Process. Lett.*, vol. 24, no. 3, pp. 329–333, Mar. 2017.

[6] J. Sulam, and M. Elad, "Large inpainting of face images with trainlets," *IEEE Signal Process. Lett.*, vol. 23, no. 12, pp. 1839–1843, Dec. 2016.

[7] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4681–4690.

[8] Y. Wang *et al.*, "Wide-context semantic image extrapolation," in *Proc. IEEE Conf. Comp. Vis. Pattern Recognit.*, 2019, pp. 1399–1408.

[9] M. Sabini, and G. Rusak, "Painting outside the box: Image outpainting with GANs," 2018, *arXiv:1808.08483*.

[10] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*.

[11] L. Xu *et al.*, "Structure extraction from texture via relative total variation," *ACM Trans. Graph.*, vol. 31, no. 6, pp. 1–10, 2012.

[12] Y. Ren *et al.*, "StructureFlow: Image inpainting via structure-aware appearance flow," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 181–190.

[13] L. Xu *et al.*, "Structure extraction from texture via relative total variation," *ACM Trans. Graph.*, vol. 31, no. 6, pp. 1–10, 2012.

[14] L. Xu *et al.*, "Image smoothing via L_0 gradient minimization," in *Proc. 2011 SIGGRAPH Asia Conf.*, 2011, pp. 1–12.

[15] J. Y. Zhu *et al.*, "Toward multimodal image-to-image translation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 465–476.

[16] S. Iizuka, "Globally and locally consistent image completion," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–14, 2017.

[17] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[18] M. Arjovsky, C. Soumith, and B. Lon, "Wasserstein GAN," 2017, *arXiv:1701.07875*.

[19] I. Gulrajani *et al.*, "Improved training of Wasserstein GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5767–5777.

[20] M. Norouzi, D. J. Fleet, and R. R. Salakhutdinov, "Hamming distance metric learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1061–1069.

[21] I. Melekhov, J. Kannala, and E. Rahtu, "Siamese network features for image matching," in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2016, pp. 378–383.

[22] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2414–2423.

[23] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3213–3223.

[24] Z. Yang *et al.*, "Very long natural scenery image prediction by outpainting," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 10561–10570.

[25] D. P. Kingma, and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[26] K. Nazeri *et al.*, "EdgeConnect: Generative image inpainting with adversarial edge learning," 2019, *arXiv:1901.00212*.